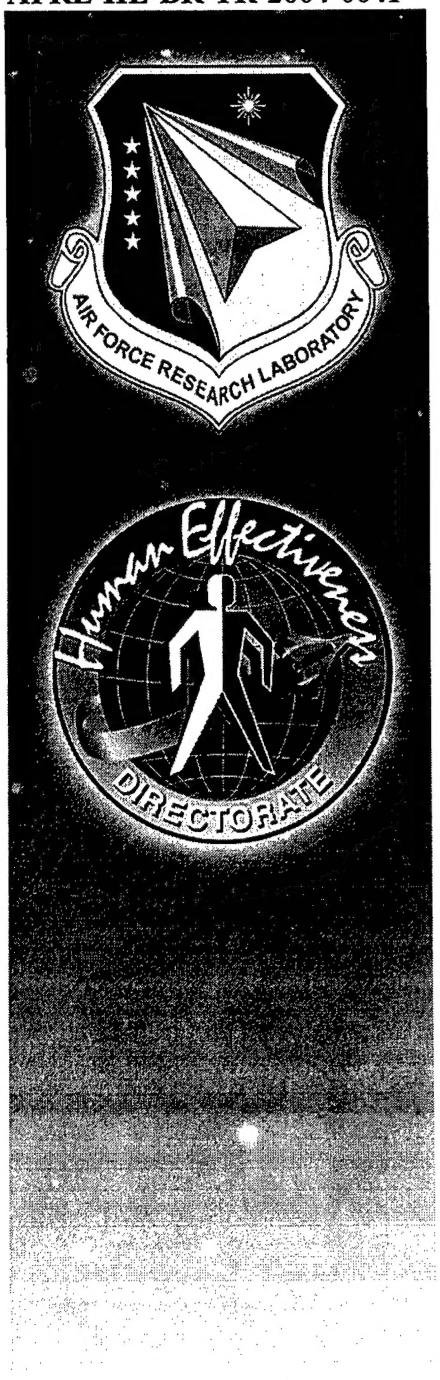
AFRL-HE-BR-TR-2004-0041



United States Air Force Research Laboratory

DO TEAMS ADAPT TO FATIGUE IN A SYNTHETIC C2 TASK?

Scott Chaiken
Lt. Christopher Barnes
Donald Harville
James C. Miller

HUMAN EFFECTIVENESS DIRECTORATE
BIOSCIENCES AND PROTECTION DIVISION
FATIGUE COUNTERMEASURES BRANCH
2485 GILLINGHAM DRIVE
BROOKS CITY-BASE TX 78235

Linda Elliott

ARMY RESEARCH LABORATORY
USAIC-HRED FIELD ELEMENT
FT. BENNING, GA 31905-5400

Mathieu Dalrymple
Phillip Tessier
Joseph Fischer

GENERAL DYNAMICS COMPANY
2485 GILLINGHAM DRIVE
BROOKS CITY-BASE TX 78235

Cory Welch

NTI, INC.
2485 GILLINGHAM DRIVE
BROOKS CITY-BASE, TX 78235

May 2004

NOTICES

This report is published in the interest of scientific and technical information exchange and does not constitute approval or disapproval of its ideas or findings.

This report is published as received and has not been edited by the publication staff of the Air Force Research Laboratory.

Using Government drawings, specifications, or other data included in this document for any purpose other than Government-related procurement does not in any way obligate the US Government. The fact that the Government formulated or supplied the drawings, specifications, or other data, does not license the holder or any other person or corporation, or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

The Office of Public Affairs has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

//SIGNED//
SCOTT R. CHAIKEN
Project Scientist

//SIGNED//
F. WESLEY BAUMGARDNER, Ph.D.
Deputy, Biosciences and Protection Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
May 2004	Interim	Feb 2002-Feb 2004

4. TITLE AND SUBTITLE
Do Teams Adapt to Fatigue in a Synthetic C2 Task?

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
Chaiken , Scott, Elliott, Linda, Barnes, Christopher, Harville, Donald, Miller, James C., Dalrymple, Mathieu, Tessier, Phillip, Fischer, Joseph, Welch, Cory

5d. PROJECT NUMBER
7757

5e. TASK NUMBER
P9

5f. WORK UNIT NUMBER
04

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Army Research Laboratory
USAIC-HRED Field Element
Ft. Benning, GA 31905-5400

General Dynamics
2485 Gillingham Drive
Brooks City-Base, TX 78235

NTI, Inc.
2485 Gillingham Dr
Brooks City-Base, TX 78235

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Human Effectiveness Directorate
Biodynamics and Protection Division
Fatigue Countermeasures Branch
2485Gillingham Drive
Brooks City-Base, TX 78235

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/HE

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-HE-BR-TR-2004-0041

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release, distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
There has been little systematic research on fatigue for teams when compared to individuals. We investigated how team performance degrades with sustained operations on a PC-based moderate fidelity air battle management synthetic task. Teams of ISR, Strike, and Sweep battle managers conducted 8 one-hour missions from 1830 to 1030 the following day, along with performance assessment batteries (during alternate hours of testing). This modest fatigue protocol allowed us to explore team fatigue assessment for both mission outcome and team process, complementing past analyses (Elliott, Coovert, Barnes, Miller, 2003; Harville, Elliott, Coovert, Barnes, Miller, 2003). In addition, one of the team roles had lower workload, allowing us to assess whether the lighter role would receive greater workload in fatigued vs. non-fatigued conditions, as a team-adaptive fatigue countermeasure. Our results showed participants performing more poorly while fatigued both on cognitive tests and on one dimension of mission outcome (number of enemy kills) but not on others (friendly losses to fuel outs and hostile actions). General activity level for the team roles declined with fatigue (number of orders issued, information seeking). Finally, while roles recognized the value of offloading work onto the lighter role, this tendency did not significantly increase with fatigue.

15. SUBJECT TERMS
Fatigue, Teams, Command and Control, Team Performance, Synthetic Task Environments

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Scott Chaiken
Unclass	Unclass	Unclass	Unclass	24	19b. TELEPHONE NUMBER *(include area code)* (210) 536-2870

CONTENTS

FIGURES

TABLES

Abstract

There has been little systematic research on fatigue for teams when compared to individuals. We investigated how team performance degrades with sustained operations on a PC-based moderate fidelity air battle management synthetic task. Teams of ISR, Strike, and Sweep battle managers conducted 8 one-hour missions from 1830 to 1030 the following day, along with performance assessment batteries (during alternate hours of testing). This modest fatigue protocol allowed us to explore team fatigue assessment for both mission outcome and team process, complementing past analyses (Elliott, Coovert, Barnes, Miller, 2003; Harville, Elliott, Coovert, Barnes, Miller, 2003). In addition, one of the team roles had lower workload, allowing us to assess whether the lighter role would receive greater workload in fatigued vs. non-fatigued conditions, as a team-adaptive fatigue countermeasure. Our results showed participants performing more poorly while fatigued both on cognitive tests and on one dimension of mission outcome (number of enemy kills) but not on others (friendly losses to fuel outs and hostile actions). General activity level for the team roles declined with fatigue (number of orders issued, information seeking). Finally, while roles recognized the value of offloading work onto the lighter role, this tendency did not significantly increase with fatigue.

Introduction

Teams are an ubiquitous part of modern warfare, and given highly specialized and complex weapon systems such teams tend to be functionally organized. A valid scientific question, with practical applications, is how the performance of teams differs from the performance of individuals and/or how much the individual performance paradigm differs in kind from a hypothetical "team" performance paradigm (c.f. Zalesny, Salas, & Prince 1995). In the case of our particular study, we further complicate this basic question by considering performance and stressors, namely fatigue. This sets up a two-pronged challenge: 1) assessment of team functioning and 2) assessment of change in team functioning under fatigue. A little reflection suggests that these two issues are completely dependent; ones ability to measure change in team process depends on the validity and completeness of team performance metrics that are currently available. The team-performance metric field is by no means a well-understood field, primarily because of the vast set of contexts in which term the "team" applies. Though frameworks for how to understand team decision processes (e.g. Hinsz, 2001; Hollenbeck, Ilgen, Sego, Hedlund, Major, & Phillips, 1995) and taxonomies for team interaction (e.g. Klein, 2001) have been advanced, their application to team measurement is sparse in realistic settings.

Unlike team processes, fatigue processes are fairly easy to measure and induce in an individual. Additionally, the relationship between fatigue and cognitive efficiency in the individual has been studied and there are well-articulated models for describing task performance under varying levels of fatigue (e.g. Hursh, 1998). Such models predict fatigue effects on one-dimensional tasks, usually simple tests. Inference to a complex job-situated task may be straightforward, if an *individual* does the job-situated task and that task relates to others that have been studied under fatigue protocols. For instance, a computerized psychomotor task, which can be studied under fatigue, can allow us to conclude something useful about fatigue and the probability of mishap in air-to-air refueling. In this particular paper we seek analogous methods for understanding the consequences of fatigue in team tasks, specifically drawn from the Command and Control (C2) domain.

In contrast to models of how fatigue effects individual performance, models for team performance under fatigue are less well articulated and possibly do not exist (at least we are unaware of them). There may be some agreement that fatigue, as a type of stressor, could deconstruct "team-ness", via a focusing in on tasks more related to ones respective role and less related to tasks of ones teammates (Endsley, 1999; Klein, 1996). On the other hand given the highly specialized nature of teams in modern warfare, this doesn't necessarily imply fatigue will affect team performance. For instance, teams can be designed so coordination between team members is minimized or automated, and this is generally a good design heuristic (MacMillan, Paley, Levchuk, Entin, Freeman, Serfaty, 2001).

However, team coordination demands may also be high and/or team organization may not be strictly functional. In this case, there is an obvious ambiguity for extrapolating individual fatigue effects to fatigue effects for teams engaged in an overall goal. If team members can share responsibilities with some flexibility, do teams have strategies available (that individuals don't) to compensate for their individual fatigue, for example pooling their efforts or redistributing their workload on some tasks?

In this study we explore issues such as these using a Synthetic Task Environment of the Airborne Warning and Control System (AWACS) cell. This cell is evolving operationally to include Intelligence, Surveillance, Reconnaissance (ISR) functions (e.g. the E-10A MC2A), in addition to responding to ground threats -- a Strike role -- and responding to air threats – a "defensive counter air" or Sweep role. The platform we considered allowed us to observe these three roles in an AWACS C2 team task, both fresh and fatigued.

Figure 1 shows a notional (but accurate) schematic of the team task we studied, using one of the scenario types we employed (i.e. SEAD, see Procedure). As a schematic, it is understood that any given part of the diagram might be behaviorally measured in a number of different ways (it is also understood that the diagram is not complete). What are depicted are the 3 roles (colored bubbles), each having a set of nearby tasks (clear bubbles, usually labeled). Arrows going into a task bubble reflect an influence on how well the task is executed. Arrows going out are influences on other tasks. Influences on a task can come both from the individual role charged with that task and from other tasks in other roles. Individual task outcomes affect a mission outcome, which is some bottom-line (and not necessarily homogeneous) measure of how well the team actually did. Other particulars of the diagram are worth noting, such as the ISR role having less work to do than the other roles. This was in fact a design feature of the study.

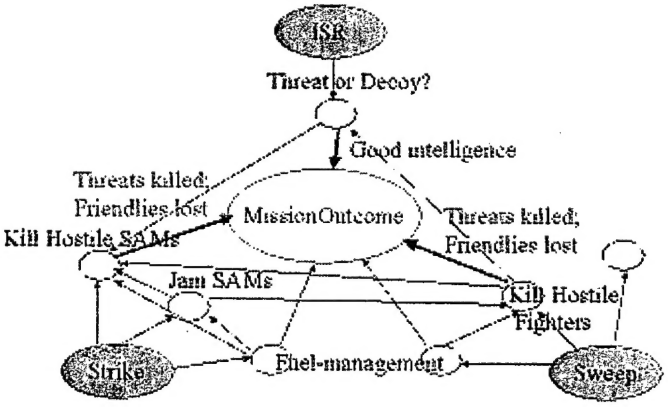


Figure 1: Execution Phase of an AEDGE mission

Our study is part of a series of studies that investigate fatigue measurement within a team performance paradigm. Other studies on the same team corpus data include analyses of verbal communications recorded during mission execution (Harville, Elliott, Coovert, Barnes, and Miller 2003) and analyses of physiological, personality, and abilities measures on team performance via multi-level modeling (Barnes, Elliott, Coovert, Harville, 2004). We add that other kinds of data have been collected on this corpus (e.g. proctor ratings on behaviorally anchored scales) and have not yet been fully analyzed. These are all part of an on-going exploratory process to discover what multi-level / multi-method measurement processes best allow us to assess team performance and process. The overall goal is to have better-articulated models of how to counteract fatigue in operationally relevant team settings. Fatigue countermeasures might be motivational interventions, perhaps implemented by pharmacology,

and/or cognitive-demand interventions, perhaps implemented by task interfaces (Barnes, Elliott, Coovert, and Harville, 2004).

AWACS- Agent Enabled Decision Guide Environment (AEDGE™)[1]

The current paper concentrates on that part of the team measurement that could be automatically logged within the context of the mission simulation. In many ways, this study just explores the logging potential of one particular Synthetic Task Environment (STE), for reconstituting a team's thinking in a C2 domain. While this sounds like an applied and uninteresting reduction of the goals above, it is important to keep in mind a particular fact of scientific life. Our measurement instruments greatly constrain the adequacy of our theories. Or as Anderson, (1987, p. 479) puts it "Much of the history of scientific progress seems to depend on the availability of tools that raise the informational yield associated with an experiment". Therefore the assessment of such loggings, in terms of their ability to support theory building is an important primary activity. Such automatic loggings (i.e. how best to log team performance data) are surely an important issue for developing new concepts of operations and evaluating simulation trainers in team contexts. That is, such loggings would seem to hold the most promise for being able to objectively, rapidly, and reliably assess team (or individual) performance after the fact, and perhaps, redress team (or individual) performance during mission execution.

We believe STEs are potential upgrades to typical laboratory studies investigating team or group processes. The influence of these tasks is gaining position in studies of group knowledge and group dynamics (e.g. Cooke, Kiekle, Salas, Bowers, Stout, Cannon-Bowers, 2003; Porter, Hollenbeck, Ilgen, Ellis, West, Moon, 2003), and we have argued, as others, that increased realism does not necessarily lead to decreased experimental control (Elliott, Dalrymple, Schiflett, Miller, in press; Brehmer & Dorner, 1993).

We were lucky to be able to beta-test the AWACS-AEDGE (hereafter AEDGE, for brevity) in our investigations of team fatigue. We think our goals will be partially realized given the history of this platform, which originated as a research instrument and later became a trainer and a war-games interface (see www.21csi.com for details). The AWACS version of the AEDGE is based on cognitive and functional analysis of C2 mission, tactics, team member roles, and role interdependencies (Barnes, Petrov, Elliott, & Stoyen, 2002). Previous studies with earlier versions of this STE have mostly involved evaluation of its Decision Support System (Chaiken, Elliott, Dalrymple et. al. 2001) and naturalistic decision-making during fatigue (Elliott, Barnes, Brown, et. al., 2002).

Our particular version logged significant mission scenario events (losses, kills) and operator actions (orders, information-seeking, scope navigation). These are in sufficient detail (e.g. assets involved, timestamps) so we could attempt to reconstitute team process for missions, and try to detect differences under both fresh and fatigued conditions. More discussion of this is given in the context of the results. At the end of our study, we hope to have a clearer conception of how team performance was impacted by fatigue and what adaptive strategies of teamwork may have been naturally employed to counter them. We investigate these issues using repeated-testing on similar but non-identical mission scenarios, carried out both in a (relatively) fresh and (more)

[1] In the research literature with this product, the acronym AEDGE usually has the word "Group" instead of "Guide". Here we use "Guide" to balance things out and shift to the dominant meaning the acronym has since taken on.

fatigued state. Few constraints are put on the teamwork process other than an initial self-assignment of participants to roles and an initial assignment of assets.

Method

Participants

Research participants were an unbiased sample from the pool of USAF officers awaiting Air Battle Management Training at Tyndall Air Force Base. A total of ten 3-person teams participated in this study (23.3% female, mean age 26, standard deviation 3.1 years). Participants were chosen simply on the basis of their availability at the time of testing. These participants had completed the Aerospace Basics Course, which provided background doctrinal knowledge; however, no participant had significant field or simulation experience with Air Battle Management.

General Procedure

Each participant had a 40-hour (5 day) training and preparation period for the fatigue protocol. This period included one hour of administrative processing, nine hours of *practice* on the Automated Neuropsychological Assessment Metric (ANAM, a cognitive test battery; Reeves, Winter, Kane, Elsmore, & Bleiburg, 2001), and 30 hours of gradated training on Command and Control interface functions for the AEDGE™ AWACS simulation. The AEDGE (Petrov and Stoyen, 2000) is a combination Airborne Warning and Control Systems (AWACS) simulation and decision support system. It can be used in either mode, and in the current study we used the simulation aspects of it. AEDGE training included briefing presentation slides on AWACS control functions and tactics in the AEDGE context, non-team practice on the AEDGE simulation platform[2], and at least 3 team plays, with each participant experiencing each role under conditions similar to the experimental sessions. Just prior to the experimental sessions subjects self-assigned which role they would play throughout the rest of the protocol. However, the flexibility of each role was also stressed during training by practicing the asset "handover" functions of the AEDGE.

Using their preferred ergonomics (e.g. chairs, monitor placement), participants began the fatigue protocol at 1830 on the last day of training (Friday) and ended testing at 1100 the following morning. They played new AEDGE scenarios during odd hours of the protocol and did ANAM and other testing during the even hours. AEDGE testing sessions were eight 40-minute mission executions. Twenty additional minutes of AEDGE time were used for mission planning (constrained at 10 minutes before simulation start) and mission debriefing (up to ten minutes after mission execution, but mostly used for proctor-monitored breaks, data-collection, and prepping for the next session).

Our version of the ANAM tests used 4 simple tasks assessing simple reaction time, working memory, mathematical processing, and spatial processing. As with the AEDGE, ANAM administration took less than an hour. After each ANAM session, physiological data (e.g., temperature), self-report data (mood-state, and sleepiness scales), and other experimental tasks

[2] As the AEDGE is highly agent enabled, these sessions involved practice in the context of automated enemies and teammates.

(i.e. a multi-tasking test) were collected for exploratory reasons. ANAM filler activities are not further discussed in this paper.

AEDGE Materials and Procedure Details

Mathieu Dalrymple, a former Weapons Director Instructor, wrote the tactical scenarios for the current study. These were developed to capture core team coordination and problem solving in the AWACS domain. Intelligent agent technology controlled the actions of the enemy and provided a realistic and proactive enemy. Use of agent technology in this way serves both to increase experimental control and to maintain high operational relevance (Elliott, Dalrymple, Schiflett, & Miller, in press).

Three distinct C2 functional roles were trained for AEDGE scenarios: 1) an ISR role, owning surveillance aerial vehicles (UAVs and other ISR assets), 2) a Strike role owning air-to-ground bombers (some with air-to-air capability) and airborne jammers, and 3) a Sweep role owning air-to-air fighters. In addition, a fourth role, "High Value Asset" (HVA), controlled Tankers on programmed routes and friendly SAM sites. HVA was completely agent-controlled. However, its assets (e.g. Tankers) could be transferred to other roles given transfer requests from these roles. Unlike the live roles, HVA could not deny a transfer request. An important feature of this simulation platform is the ability for roles to swap resources and divide the workload any way they like. The only advice/constraints given to participants for resource swapping was not to transfer assets to HVA other than returning assets HVA might have originally owned.

Mission scenarios were designed to require a demand for communication and coordination, and perhaps some adaptive problem solving. The former followed because the roles were interdependent and functionally organized (at least initially). For instance, ISR started out with nothing but Intelligence assets. ISR had to identify the real surface-to-air missile (SAM) threats as opposed to harmless decoys. Fifty-percent of the candidate SAM threats were decoys. Once ISR identified a valid threat, Strike could kill them, but Strike needed protection from enemy fighters via Sweep to do that. However, Sweep's protection of Strike also depended (somewhat) on Strike's protection of Sweep from SAM threats (e.g. using jammers against valid SAM threats). Adaptive problem solving to time-critical situations was promoted in each scenario by having a mix of a priori and pop up targets (both enemy fighters and enemy SAM sites). Pop up targets could require re-planning from missions already in progress. The predictability of pop up targets was manipulated as a "Study Design Factor" (see below).

Scenarios were carefully constructed to promote equivalence in task demand. In particular, scenarios had (a) the same roles, (b) equivalent assets initially allocated to each friendly role and equivalent pop up assets later in the scenario, (c) equivalent hostile threats, (d) equivalent timing and tempo of events, and finally, (e) equivalent geographic distances for hostile and friendly interaction.

Geographic distances affect the timing of hostile-friendly encounters and thus directly affect the tempo of workload demand. Changing the geography of a scenario, or its "surface" structure, was designed to reduce recognition of the underlying "deep" structure of each scenario. One scenario may be located in the geographic region of Taiwan, while another would be in Sri Lanka. The number and distribution of assets on different geographies would always be equivalent and have similar spatial distributions, but would not be spatially constant from scenario to scenario (i.e. in terms of exact pixel locations). Another way deep structure was masked was by changing the type of theatre mission from scenario to scenario. In one version of

a scenario, hostile threats were comprised of enemy surface-to-air missile sites. This situation is known as a SEAD scenario (suppression of enemy air defense). In another version, the hostile targets were theatre ballistic missiles (typically referred to as SCUDs). Finally, the third version of a scenario had hostile naval ships as enemy targets (SEA). The way these geographies and mission scenarios were distributed across fatigue was also part of a "Study Design Factor". Of the eight scenarios used, four were SEADs, two were SCUDs, and two were SEAs.

AEDGE Study Design Factor

Creating the eight AEDGE missions for the fatigue protocol involved an experimenter learning process that resulted in a "study-design factor"— 1st five teams (less optimal design) vs. 2nd five teams (better design because it was informed from analyses on the first half). What we did differently in the last half of the study was to better balance the type of mission within the early vs. late sections of the fatigue protocol. With the 2nd five teams, mission types occurred equally often early as late, with comparable mission types optimally spaced (e.g. a SCUD scenario early had a matching SCUD scenario four AEDGE testing sessions later). In contrast the first half was partially balanced (balanced on SEA and SEAD, but both SCUDs were placed at the end of the protocol). A second improvement from the first half was to make the distribution of "pop up" enemy activity less predictable throughout the entire protocol. Apparently, having temporally similar pop ups in different scenarios, enabled some subjects to predict enemy pop ups. For later groups of subjects, enemy pop-ups were smeared over wider temporal bands, but with the same average pop up time. Significant effects discussed do not interact with the Study Design factor.

Results

Individual Measures of Fatigue

One of our first analyses assessed the presence of fatigue. This is typically done using a well-established performance assessment battery on which fatigue effects have been robustly demonstrated. We intentionally chose a subset of the ANAM as one such battery. Each participant provided eight scores (one for each point in the fatigue protocol) on each of four tests: 1) a continuous performance task (CPT, the "two-back" version of a continuous recognition task using letters), 2) a math processing/math knowledge task (MATH), 3) a simple mean reaction time task (SMRT), and 4) a spatial comparison task (SPAT). Each person's score on any given test was scaled by his or her study variation for that test. For instance, each of the RTs a participant provided for SMRT had each of their 8 scores "standardized" by subtracting off the mean of their 8 RT scores and then dividing the score by the standard deviation of their 8 RT scores. These standardized scores were then averaged across subjects and across tests to give an aggregated ANAM curve shown in Figure 2. [3] The curve shows a canonical fatigue pattern of performance (i.e. initial increase, troughing in the early morning hours and recovering somewhat at the end of the protocol). The trends are significant as indicated by the error bars (one standard deviation on either side of the means).

[3] We used "throughput" scores on all tests except Simple Mean Reaction Time (SMRT). Throughput is a metric that combines speed and accuracy. This metric was not appropriate for SMRT because that test had near 100% accuracy across all subjects. Instead, only the mean reaction time was used to score that test. Also z-units for SMRT were reflected about 0 so that large RTs would indicate poorer performance.

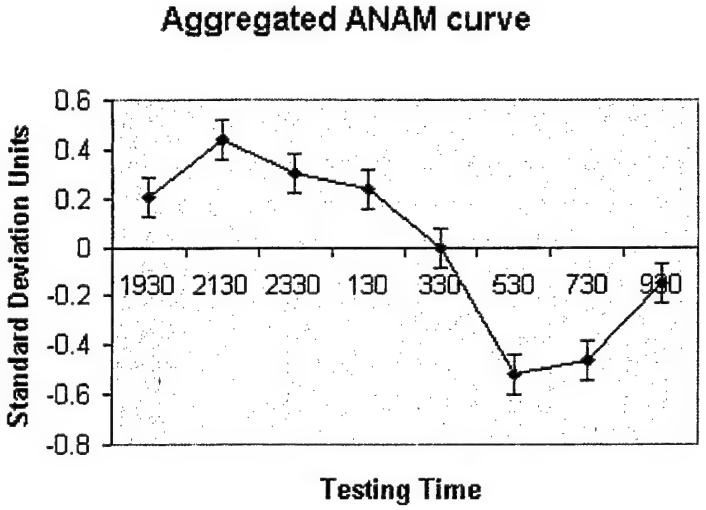


Figure 2. (See text for details).

ANAM Test	Paired *t(29)*-statistic	Correlation *r*(28)
CPT	5.41	0.62
MATH	2.89	0.77
SMRT	2.74	0.48
SPAT	3.76	0.88
SLEEPY?	15.4	0.70

Table 1. *Individual ANAM tests compared early (less fatigued) to late (more fatigued).*

In addition, the ANAM tests were divided into early (average of first 4 testing sessions, hereafter just "Early") and late trials (average of second 4 testing sessions; hereafter just "Late"). This is a data-analysis tactic that we apply later on the AEDGE measures. One can do this sort of simplifying analysis both to enhance the clarity of our results, but more importantly we seek to ameliorate suspected differences in our AEDGE scenarios that we were not aware of prior to data collection. That is, aggregating over larger comparison sets better protects us from non-equivalencies that we are not systematically in control of prior to data collection. The "Early vs. Late" results for the ANAM are shown in Table 1. The table shows the paired t-test statistic (n=30) comparing early and late average scores as well as the correlation between the scores. Dividing the ANAM data into Early and Late segments still allows us to detect fatigue effects on the ANAM tasks (e.g. all *t*s are significant in the correct direction, with some tests more sensitive to fatigue than others). In addition, Table 1 presents the Early/Late comparison of a 7-point sleepiness-rating questionnaire (also with t and r). The average rating Early vs. Late was 2.6 vs. 4.6. In summary, fatigue effects on the cognitive tests, as well as participant sleepiness ratings, indicate our participants were fatigued under this protocol.

AEDGE Mission Outcome measures[4]

Our ISR role model was less developed than the other roles, so we had no clear performance measure for the "good intelligence" influence on "mission outcome" (see Figure 1). However, the other two roles have clear metrics for outcome and these were tallied (i.e. threats killed and friendlies lost). These tallies are displayed in Figure 3 as a bar chart. Before describing these results, we describe our presentation conventions. We typically use total counts on the y-axis, with a unit of aggregation of 4 testing sessions—i.e. the 4 early ones (blue bars) and the 4 later

[4] We have complete data on mission outcome, but some missing data on individual actions. The latter leads to smaller than expected *df* on some analyses reported below. Different log files handled the mission outcome, event, and individual actions, and only the individual action files have missing data (two participants, one session). If we look at estimating missing data from extant data, results reported do not change and tend to be more significant. Except where noted we report analyses that reflect removal of participants with missing data.

ones (purple bars). Given this convention one should understand the word EARLY in the bar-chart legends to mean relatively fresh, the word LATE to mean, "fatigued". As the bars are aggregated over 4 sessions, to get a per-session indication of "Ops Tempo" or workload, divide the bar height by four. Along the x-axis we'll typically show interesting event categories. In general, these categories are compared Early to Late using teams as the unit of measurement (n=10), although we also often look at event categories for each of the 3 roles (also n=10 comparisons). We make no attempt to present everything on the same scale, but we do try to group our event-category analyses into meaningful ensembles.

Figure 3 shows no difference in fuel management outcome ("No Gas"), or friendlies killed by hostiles ("FKBH") between early and late sessions. However, the difference in hostile kills by friendlies is significant ($t(9)=2.34$, $p<.05$, 2-tailed), once the kills by friendly SAMS are subtracted out. Recall that friendly SAM sites were controlled by HVA, an agent role, which never tired. Hostiles that were killed by friendly SAMs were counted as enemy penetrations (Hpenets). The small increase (a scaling illusion) in hostile penetrations with fatigue was significant ($t(9)=3.65$, $p<.01$, 2-tailed). This is a redundant (or concordant) finding with the HKBF effect—that is the fewer enemies that are killed by the live roles, the more that can be killed by the HVA agent. Finally, correlations are of interest for this analysis. In particular, a correlation between Early and Late outcomes says something about "individual differences" among teams. Table 2 shows these correlations for the 3 event types that humans were responsible for. These are generally significant (1-tailed) and if the measures are aggregated (i.e. HKBF-FKBH-No_Gas), the correlation between early and late performance on teams is the most significant. This indicates correlation among the mission outcome measures, as one would expect (e.g., the higher friendly attrition rate going with lower enemy kill rate).

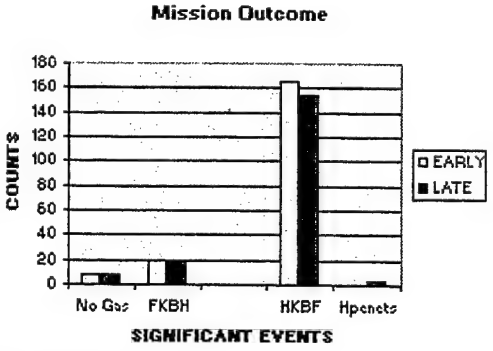


Figure 3. (See text for details).

Measure	$r(8)$	p, 1-tailed
No Gas	.55	.05
FKBH	.68	.025
HKBF	.64	.025
Aggregated	.83	.01

Table 2. *Early to Late correlations among team outcome measures. (See text for details).*

In summary, some fatigue impact on Mission Outcome is evident. In particular, fewer enemies are being killed, with no change in a team's tendency for risk aversion (i.e. friendly fuel outs and attritions are equal early to late). Moreover, our scenarios can discriminate between teams that are very good and teams that are not as good. The latter effect is important as it shows our scenarios are not trivial, otherwise they could not discriminate team ability (we'll return to this issue in the discussion).

AEDGE Individual Process Measures: Counts

Our measures of "team" process at the individual level are in their infancy. We would like measures of problem recognition, solving, and decision-making in the individual, as well as measures of team coordination while doing any of these. However, for our current loggings, which were not designed by us, we identified the best available indicators as steps toward more direct ways to measure individual action relevant to team process. The three most obvious individual measures we could get from our loggings were participant scope changes, information window openings on scope entities, and orders given by participants to their assets. Measures of scope changes were essentially content-free, logging only their occurrence. Measures of information-window openings had a content reflected by the type of entity inquired about. Measures of orders were multi-dimensional, having both a category for the order and some content (subject and object of the order). Orders are substantially more complex and so are given a more detailed analysis.

Measures of scope changes (frequency of scope adjustments Early vs. Late) did not yield any significant effects. Our loggings did not enable us to differentiate between the types of scope adjustments (i.e. zooming in, zooming out, and panning). The only thing worth noting about scope adjustments is that they were reliable across individuals. In other words, some people like to adjust their scope frequently, while fresh or fatigued, while some people don't. This held true mainly for the ISR and Strike roles (Sweep being less reliable in this regard). The average within-role correlation is $r(8)=.78$, range .43.[5]

Figure 4 displays results for Information Window openings. Three categories were easily assessed from the loggings, Information Windows on ones own assets (MINE), ones friends (i.e. assets owned by the other roles, or FRIENDS), and on hostiles (ENEMY). Information given on own assets was the most complete information (e.g. exact fuel and armament); information given on other types of assets was more or less restricted to what was available from the scope (e.g. bearing, speed, altitude), and what might be intelligently guessed. From the point of view of the task, there was very little reason to open up an information window on a friend or enemy asset. For one thing the decision-aid aspects of the platform provided a "threat-ring" capability, which performed some of the functions of the information window, at least with respect to remaining armament (apparently armament firings on both friendly and enemy sides were tallied by the platform). However, for ones own assets, the information window was the only way to get current fuel. In fact, information window openings, MINE, was the most numerous category. In addition the overall drop in information requests from Early to Late was significant within the MINE category ($F(1,25)=7.64$, $p<.011$). The apparent interaction of role and fatigue in the MINE category (ISR stays level, the other roles drop) was not significant ($F(2,25)=1.95$) but suggestive (more on ISR later in Figure 5 and Figure 6). Again, there was a fair amount of reliability on this individual behavior (average within-role correlation, early to late is $r(8)=.88$, range = .17).

---

[5] Why give the average of the within-role correlations? We know from other analyses that the frequency of individual behaviors are driven by the role you are in, not exclusively who you are. For instance, ISR gives fewer orders per session than the other roles (e.g. ISR issues about 28% of the orders per session, with the other roles each issuing about 37%). This means a correlation across all 30 participants (i.e. across roles) includes role variation as well as the individual variation in the participant. As we are more interested in the *individual* tendency to issue orders regardless of role, the average of three ($n=10$) early-to-late correlations, within each role, removes role variation. In general, plots of the correlations were always inspected for outlier effects, and in general there are none on reported effects. Later in the paper we show a sample small-n high correlation to demonstrate how they look.

For orders, we first aggregated over all order-types within a session and did a paired t-test comparing the number of orders given in the Early sessions as opposed to Late sessions. There was an overall decrease in the amount of orders later ($t(27)=4.19, p<.01$), when each participant was compared to him or herself Early vs. Late. As with the other individual process measures, the tendency to give orders had reliability as an individual characteristic (average within-role $r(8)=.86$, range=.20).

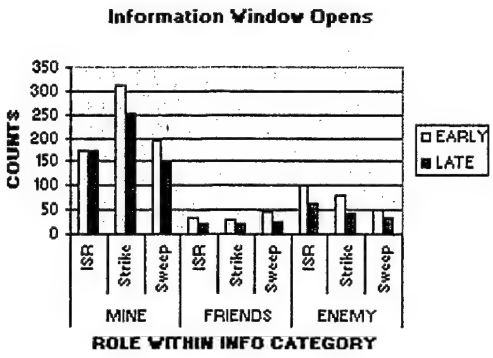


Figure 4. (See text for details).

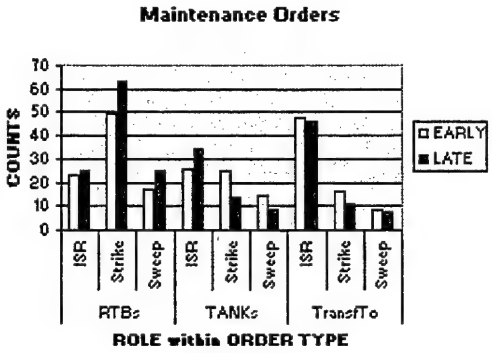


Figure 5. (See text for details).

For orders, Role by Order Type is a reasonable first cut for investigating team process and assessing differences in team process with fatigue. Within the order type category it seemed reasonable to break the analyses out by "Maintenance Orders" and "Tactical Orders". Maintenance orders are return to base (RTB) orders, orders for air-to-air refuel (TANK), and transfer of assets to another role (TransfTo, i.e. a count of transfers to each of the roles). Figure 5 displays results for maintenance orders. Each type of maintenance order has some significant effects associated with it. Firstly, the only type of order that *generally* increased with fatigue was Return To Base ($F(1,25)=4.88, p<.05$) with Strike doing more of these than other roles: $F(2,25) = 7.62, p<.05$. Strike was the role with the most initial assets. There are several different interpretations of the RTB order. One interpretation is a way to limit the workload and concentrate on a smaller playing field of assets (i.e. returning to base means "retiring" an asset). This would be reflected by those RTBs that had an asset return to base and then not be used for the remainder of the session. On the other hand bases could be more frequently used as refueling/re-armament nodes late rather than early, and this might be reflected by multiple RTBs on some assets. When we divided the RTBs, on any given asset, into RTBs once-only vs. multiple RTBs, both categories occurred often enough to contribute to the overall effect (i.e. neither alone is significant; although once-only comes closest, $t=1.86, p<.08$, 2-tailed, on a paired t-test, with, $n=28$).

TANK orders increased with fatigue *specifically* for the ISR role, but for other roles the frequency of TANK orders decreased with fatigue (for the interaction: $F(2,25)=3.65, p<.05$). This is a kind of team adaptive response to fatigue. Apparently ISR was given more of the refueling responsibility Late rather than Early. Finally, consistent with the last effect the ISR role had the most transfers to it ($F(2,25)=23.12, p<.001$). However, somewhat *inconsistent* with the

last effect, the tendency to transfer assets specifically to ISR does not increase with fatigue.[6] We look at transfer activity in more detail next.

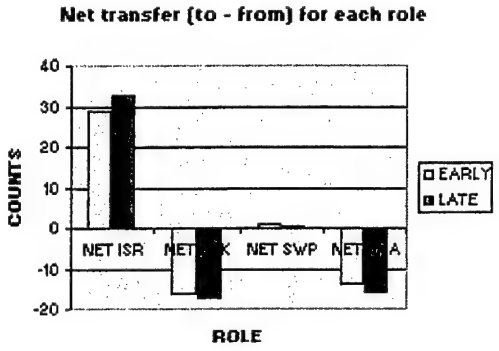


Figure 6. (See text for details).

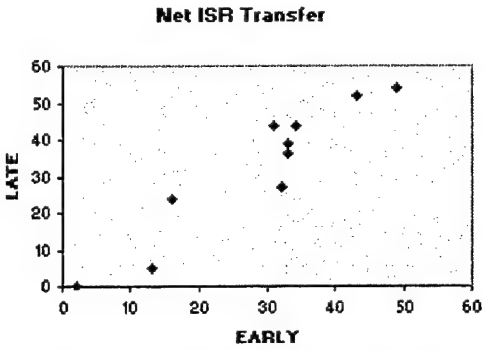


Figure 7. (See text for details).

In order to better test the hypothesis that teams were dividing the workload differently Early vs. Late we performed two analyses to explicitly address this. The first analysis looked at "Net transfer" (transfers to the role minus transfers from the role). Net transfer reflects *change* in role responsibility by the end of a session. A positive net transfer reflects gaining responsibility or workload, whereas a negative net transfer means shedding workload or responsibility. Change in role responsibility can be further assessed under different fatigue levels. With regard to a general workload redistribution hypothesis (i.e. fatigued teams will distribute the workload more equitably), one might expect ISR's workload to significantly increase in fatigued sessions as a means to compensate team fatigue overall. Hence, we expect ISR's net transfer to be positive and greater in fatigued sessions. Figure 6 shows the results of this analysis both for the critical ISR role and the other roles. While the trend is occurring, the result is not significant ($t(9)=1.79$, $p<.12$, 2-tailed). It should be pointed out that the overall size of the net transfer function supports the idea that the team (especially the Strike role) did view ISR as a role to more equitably distribute the workload on. However, this perception doesn't vary much with fatigue. Also available from the paired t-test analysis of ISR net-transfer was the correlation between Early and Late net transfer to ISR. This can be interpreted as reliability for the "team doctrine" on how much the ISR role should "pick up the slack". This correlation is $r(8)=.945$ is very high ($p<.01$). The scatter plot is shown in Figure 7. A large degree of team-individual difference on this indicates some teams utilized the ISR role significantly more than others.

Another way to check whether teams are redistributing their workload Early vs. Late is to look at order counts for a role during a session as the measure of the workload for that role during the session. As orders decrease in frequency from Early to Late, this metric could be done as a proportion (i.e. the number of orders a role gives in a session, divided by the total number of orders of that session). One can then look at whether the role workloads tend toward equality more in the fatigued sessions. This might be indexed by a decrease in the within-session standard deviation of the (proportional) workloads as one goes from Early to Late sessions (i.e. each of the roles should be moving toward 33% of the orders). This analysis yielded results similar to

[6] There was an interaction with the Study Design factor and Fatigue, such that transfers To ISR and Strike declined more with fatigue for teams in the later design. As this doesn't show up in the net ISR transfer analysis, we suspect this is not an important moderator.

Figure 6 and 7, in the sense that the standard deviation of role workloads doesn't decrease significantly with fatigue, but yet there is a very high correlation between how workload is distributed, on this metric, Early to Late among the teams ($r(8)=.97$). Figures 14 and 15 display results related to these analyses, but these are presented in the Discussion section as summary points, along with more discussion on the two types of workload-distribution descriptions.

Finally, we looked at tactical orders. These were defined as GO orders and TARGETS against specific enemies. TARGETS reflect a bombing, defensive counter air, or jamming mission (i.e. each kind of mission was logged as a TARGET order). Figure 8 shows the counts of these kinds of orders for each role. Tactical orders were emitted less often during the Late sessions ($F(1,25)=4.40$, $p<.05$; $F(1,25)=14.48$, $p<.001$, for TARGETs and GOs, respectively). One can also see in Figure 8 that ISR gave relatively few target orders relative to the other roles ($F(2,25)=25.66$, $p<.001$).

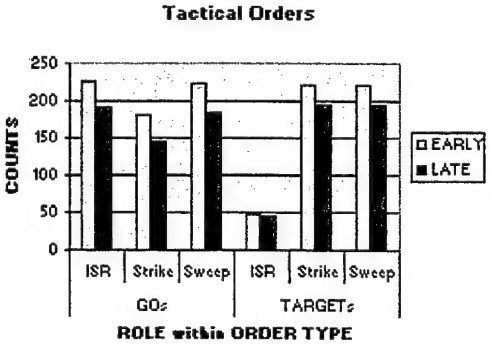


Figure 8. (See text for details).

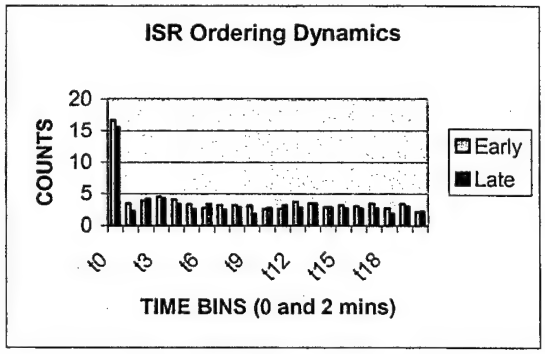


Figure 9. (See text for details).

AEDGE Individual Process Measures: Latencies

Latencies are important for discriminating between theories for why activity levels decrease in Late vs. Early sessions. Depending on the study context, one could argue either fatigue or learning (e.g. practice) effects. For instance, fewer orders emitted in Late sessions could indicate better knowledge of what orders are needed and what orders are not, rather than a reduced speed at issuing orders. To some extent our mission outcome results go against a general learning interpretation for depressed activity levels. That fewer enemies are being killed Late implies efficiency is not the reason orders are less numerous Late. Still if your theory (both scientific or just based on common sense) says that team and individual processes are slowing down, it would be nice to have latency results to back that up. In this section, we investigate possible methodologies for examining latency effects.

We looked at latency in two different ways. In a highly specific test, we considered how fast friendly assets that arrived midway through each scenario were noticed (as indexed by the latency of a first order on them). This analysis yielded no significant results (i.e. a non-significant speed up, from Early to Late). The 6 friendly "pop ups" always occurred at the same time (i.e. a 6 minute time band centered mid scenario), and their specific arrival times were always noted in the "fragmentary orders" that C2 teams studied prior to the scenario start. There

was no indication in our study that subjects forgot to utilize these resources (or that they ignored the "frags").

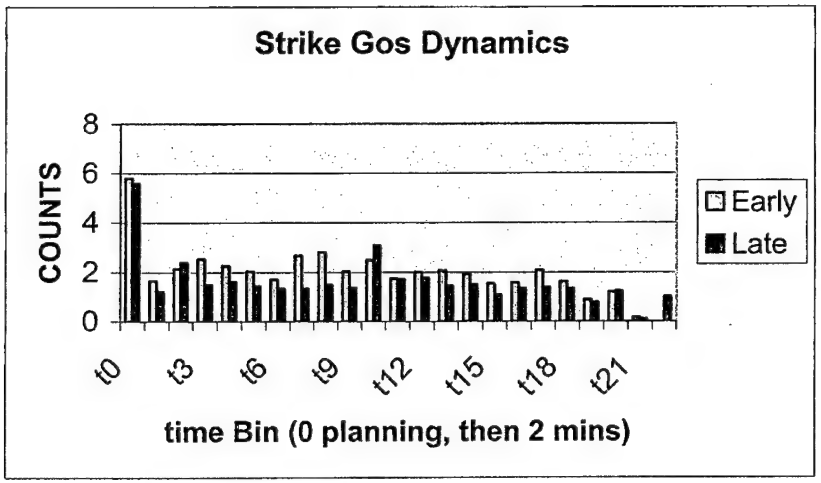
The second way we looked at latency was with latency histograms over fine time intervals, both within roles and between different order types. These results are notional, because statistical tests would depend on choosing segments of data to compare (in an ad hoc fashion). Still we can investigate the method to see if it shows promise. For all the histograms we show, the first "time bin", t0, is special. This was a time period of 10 minutes during which the simulation clock had not yet started, and during which the three roles could prepare for the mission (e.g. adjust the scope or talk among themselves). In addition, if roles had assets visible on the screen, they could pre-program sequences of orders for them. Hence, the large spike in orders counted at t0 reflects a longer time-period for that bin, as well as the ability to pre-plan missions based on the placement of enemy threats and owned assets seen at time 0. Following time 0 (t0), time bins last two minutes each (e.g. t1 being the first 2 minutes of a mission once it has started).

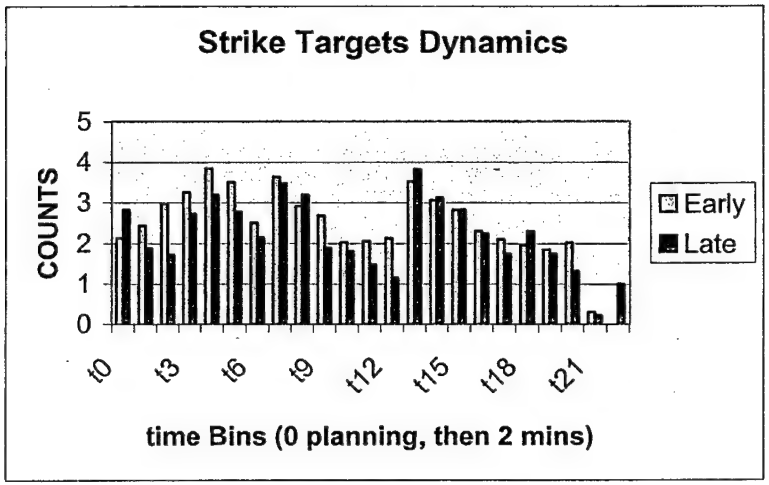
Figure 9 shows a latency histogram for the ISR role where the counts are over all order types (recall that ISR mostly issued GO orders). What's notable in Figure 9 is the lack of a fatigue effect on the mission-planning component. This may be due to the automated nature of the ISR plans. Another thing to note is the way the blue and purple bars (not counting t0) tend to follow each other (i.e. are correlated across early and late measurements). This will be even more apparent in the other roles.

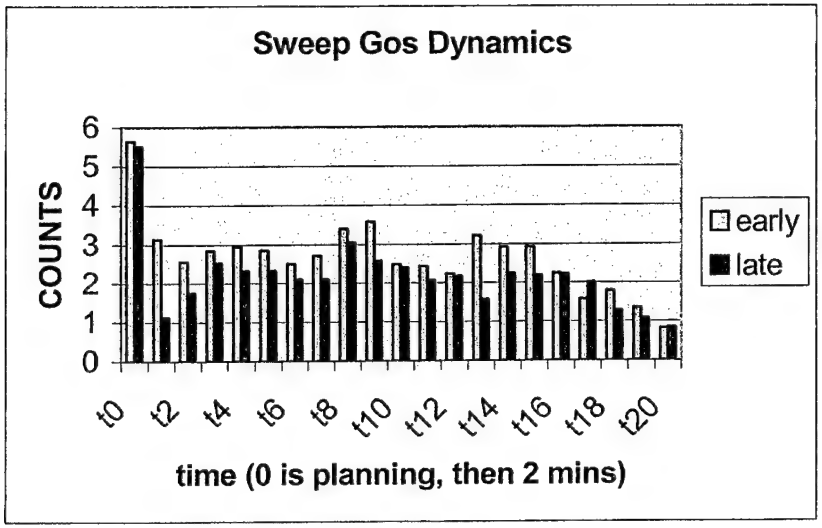
For Strike and Sweep we broke the orders out by GOs and TARGETS, as each intuitively has a latency component. GOs might indicate proactive planning, or they might indicate a "reaction time" for following the mission path of a Strike package. Targets are like a "reaction time" to a popup enemy stimulus.

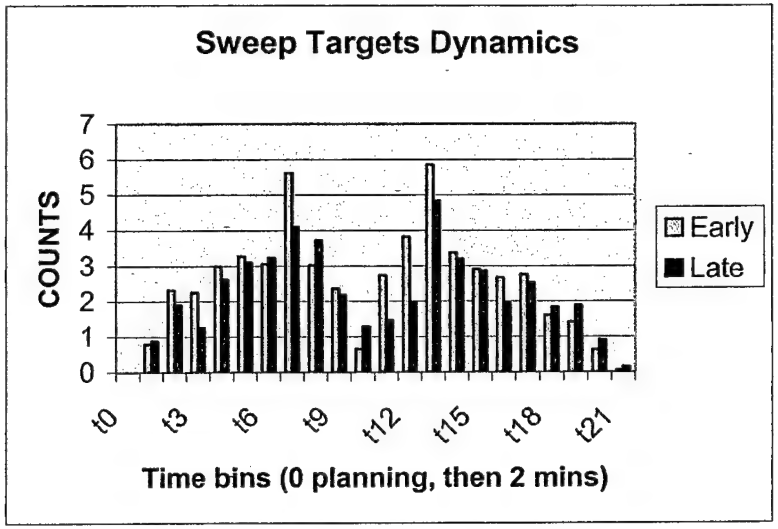
For Strike, the Go dynamics (Figure 10) were not as interesting as the Targets dynamics (Figure 11). When the scenario starts there are more target orders issued Early rather than Late (with the exception of t0), until about mid session. In the second half of scenarios, Late target counts (i.e. purple bar height) seem to catch up with Early target counts (blue bar height). As the number of targets within each session is constant, one could expect this trend. That is, one could expect relatively fresh teams to do the work faster, leading to more target orders early in the session (relative to the same teams fatigued). This could also imply a diminishing number of target orders for fresh teams later in the session, because more of the work was done early on. For the same reason, the same teams, when more fatigued, could be expected to increase target orders later on (i.e. to catch up from less Target order activity earlier on). The (admittedly slight) terminal purple spike at the end of the scenario is consistent with this interpretation.

For Sweep, the Go dynamics were interesting (Figure 12). Starting with near equal planning segments, the first moments of the scenario are marked by fresh teams seeming to reach an asymptote immediately on their level of go-activity, while the same teams, in a more fatigued state, seem to ramp up to a lower asymptote. This pattern may replicate later (t12 – t16). The Targets dynamic (Figure 13) indicates clearly two waves of enemy pop ups present in each scenario. Here the latency effects could be reflected in the slope of the blue bars vs. the purple bars for the two principle ramp ups or waves of enemy fighters.









Figures 10 – 11 (top row) and Figures 12 – 13 (bottom row). Note: The Y-AXIS, "COUNTS", on these figures (as well as Figure 9) is an ***average*** *count for early vs. late sessions; therefore to get these figures in the same metric as other figures of the paper, multiply bar height by four. See text for details.*

In summary, while the stories we describe from latency histograms are probably not statistically significant, at least they are consistent with a general sluggishness interpretation of reduced activity with fatigue. In addition, the mission-planning segment was not affected by fatigue, which was surprising. This doesn't necessarily mean planning in general is not affected by fatigue, but it could mean that teams had learned well-established procedures for initially approaching each mission. The correlation between the purple and blue bars in some cases attest to the regularity of mission executions (i.e. mission equivalence of Early and Late missions.)

Discussion

Did teams adapt to fatigue?

Our study assessed team performance under fatigue given an available synthetic task. However, we also had an intriguing side issue: whether team performance contexts could have fatigue countering adaptations not possible for individuals performing individual tasks. Did we find any evidence of such adaptation? We do not consider activity depression (e.g. fewer information window openings) as adaptation, per se. Adaptation should be some kind of strategy change in performing the task. At a global level we found very little evidence of such adaptation. This is best summarized by Figures 14 and 15, which show how roles distributed "workload" in Early and Late sessions. In Figure 14, workload is defined by the number of orders a role gives in a session divided by the total number of orders in the session (i.e. a percentage workload for a role); in Figure 15 order behaviors are measured by raw counts as was done in the Results. There are differences in these two approaches. Although, one can see a slight trend in the direction we would have thought (i.e. role workload bar-heights becoming more equitable under fatigue), the results are not significant for this metric of "strategy change" (i.e. role by fatigue interaction $F(2,27)=.93$; $F(2,27)=1.28$, percentage and counts metric respectively, with missing role data estimated by extant data). Also despite what the eye may suggest, the main effect of role on number of orders issued is not significant ($F(2,27)=2.11$, $p<.14$; $F(2,27)=3.00$, $p<.07$; counts and percentage respectively). The lack of such an effect, when ISR is issuing fewer orders (on the mean), shows ISR is being conducted very differently from team to team (i.e. significant team by ISR interactions as noted before).

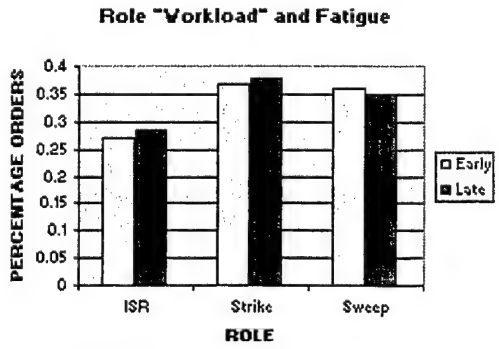


Figure 14. (See text for details).

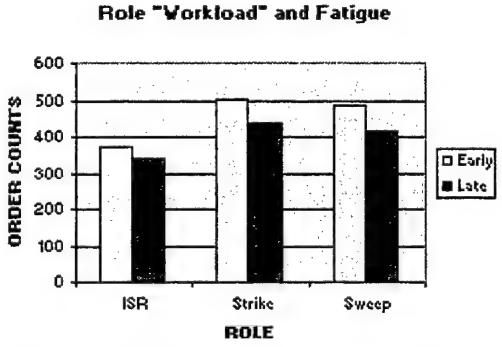


Figure 15. (See text for details).

On the other hand, at a more micro-level there were two possible indications of a strategy adaptation with fatigue. First, return to bases occurred more often in fatigued sessions. We thought this might have been a strategy to reduce workload, but found only weak evidence for this, so this is a strategy change we can't adequately explain. Second, we found that ISR's refueling activity became more important to ISR's mission for late sessions. If one were to rescale Figure 5 TANKS counts in terms of percent workload (i.e. Tank orders in a session divided by total orders in a session), TANKS does not exceed 10 percent of a role's work (on the mean) for any condition. So perhaps it's not surprising that ISR's mission change later on, occurs without a corresponding increase in the number of orders ISR issues later on. Even so, one could argue that had refueling been more prominent in our scenarios (e.g. longer missions, same number of targets), team adaptation might have been more strongly observed.

An important qualification to the statement that team adaptation *did not vary much with fatigue* is that *team adaptation does occur*. The teams recognized that the ISR role had lighter responsibilities and sought a more equitable distribution. This is indicated most clearly by Figure 6. However, teams seem to arrive at a "team doctrine" for how to view ISR role rather quickly and stick to it (i.e. as best illustrated by Figure 7). The high reliability for how teams viewed the ISR role is another indication of a lack of adaptation (i.e. lack of change in work distribution) with fatigue.

Why didn't we get stronger fatigue effects?

This is a somewhat "loaded" question given we don't know how to scale the size of the fatigue effects we actually got. It is also a complicated question given multiple dimensions of fatigue assessments exist in this task. Fatigue effects can be broken out by mission outcome measures and activity level measures, and each should be considered separately. Mission outcome was seemingly impacted only along the dimension of hostiles killed, and the effect was not visibly large. The two most important possibilities relevant to explaining this are: 1) We had no strong guarantees that learning was not continuing to occur during the later sessions (i.e. learning effects attenuate fatigue effects). 2) The friendly side was actually stronger than the enemy side as a more realistic representation of conditions under sustained operations.

Somewhat inconsistent with those possible explanations for modest fatigue effects, ceiling effects weren't generally observed. The absence of ceiling effects is demonstrated given the reasonable "individual differences" in "mission outcome ability" teams displayed (r=.83, on an aggregated measure of it). One of the virtues of a small n design is that we can actually show the individual team variation data. This is done in Table 3. If one looks at fatigue effects on the individual teams, some teams performing both poorly and fairly well seem to show sizeable fatigue effects (e.g. team 4 and team 8). One team appears to do significant learning (team 6). The best team shows hardly any effect (a true ceiling effect?).

The overall effect of looking at the individual team data is one of complexity, which may be exacerbated by possible tradeoffs among dimensions of Mission Outcome (i.e. the components in the aggregated sum). For instance, one could suppose that higher incidence of friendly attrition would *always* imply lower scores on HKBF (e.g. lack of resources to do the job), but within certain ranges of friendly attrition higher scores on HKBF might also occur (e.g. more risks taken in pursuing the enemy leading to higher HKBF). In future studies, we'd want the data

loggings to help us better assess these possibilities (e.g. detect greater risk taking vs. simply being overloaded).[7]

	EARLY					LATE			
TEAM	HKBF	No Gas	FKBH	Mission Outcome	FatigueΔ Mission Outcome	HKBF	No Gas	FKBH	Mission Outcome
4	151	12	31	108 (10)	39	123	10	44	69 (10)
6	148	13	17	118 (9)	-18	163	8	19	136 (5)
2	161	20	17	124 (8)	2	153	8	23	122(6)
9	161	7	27	127 (7)	14	137	8	16	113 (8)
7	160	9	21	130 (6)	12	145	12	15	118 (7)
10	169	2	30	137 (5)	37	135	10	25	100 (9)
1	168	15	14	139 (4)	1	164	13	13	138 (4)
5	167	3	8	156 (3)	3	172	0	19	153 (2)
8	185	5	20	160 (2)	19	166	7	18	141 (3)
3	192	1	7	184 (1)	-1	190	0	5	185 (1)

Table 3. Team variation in Mission Outcome and its Components.

Teams are sorted by Mission Outcome EARLY. Mission Outcome = HKBF-No Gas-FKBH followed by rank within EARLY/LATE in parentheses. Middle column (FatigueΔ...) gives Mission Outcome Early - Late. Other Column Headings: HKBF = Hostiles killed by friendlies; No Gas = friendlies lost to fuel management; FKBH = friendlies killed by hostiles.

Finally, we note that the best and worst performing teams are similar on how they distribute their workload as measured by the proportion-of-orders-per-role metric (i.e. average within-session sd of these proportions, collapsed Early and Late: best is .075, worst is .09). However, the two extreme teams look different on Net ISR transfer (best: 34.5; worst: 9, again collapsed across Early and Late sessions). Hence, these two ways at getting at the workload redistribution issue (i.e. team "adaptability") appear fundamentally different. Net ISR transfer is probably the better way to test a specific hypothesis that work (i.e. assets) is being off-loaded onto the ISR role. There were also teams with higher Net ISR transfer than the best-performing team, so Net ISR transfer, alone, does not explain why the best performing team is the best (i.e., $r(8) = .36$, n.s., for Net ISR transfer correlated to Mission Outcome, both collapsed on Early vs. Late).

In contrast to outcome indices, activity indices associated with thinking (information windows) and doing (orders) more noticeably decreased with fatigue. Previous analyses have shown that communication (verbal and email) on these participants also decreased with fatigue, especially for communication regarding assets and coordinating strategies. However, more essential communication about targets was less effected, so reduced communication might have been

[7] One can also explore non-linear measures of Mission Outcome, such as kill ratio (i.e. divide HKBF by the sum of No Gas and FKBH). We think this is ultimately a bad idea as it goes against intuition more often than not. For instance the best team has a kill ratio 24 vs. 38 Early to Late. This suggests a big improvement (i.e. a lot of learning) while the data appear better characterized as staying the same Early to Late. Similarly team 8 gets 7.4 vs. 6.6 on the ratio metric Early vs. Late, suggesting a modest drop in effectiveness. However, on a linear metric, the team's delta of 19 suggests a bigger drop in mission effectiveness. Finally, while a paired t-test on the Early vs. Late scores for the linear Mission Outcome aggregate is significant ($p<.05$, 1-tailed, paired-t test), the same test on team Early/Late kill ratios would not be.

attempts to reduce workload wherever possible (Harville, Elliott, Dalrymple, Barnes, Miller, & Coovert, 2003).

However, we think at least some of the automatically logged activity indices reflect bona fide fatigue effects and not learning effects (as when unnecessary or lower priority actions are deleted). Two aspects of the data support fatigue. One is the qualitative forms of the latency histograms (Figures 10 – 13) which can be said to support (or at least not refute) sluggishness in the team actions. More importantly, not all activities that drop in frequency can be a reflection of increased efficiency. In particular, target orders drop in frequency with fatigue (e.g. see Figure 11 and 13), and as the number of enemy targets in a scenario is constant, less targeting behavior can be strongly attributed to acts of omission (leading to a reduced HKBF score).

Where do we go from here?

In future studies we will be taking more direct control of both the scenario design and team performance measurements (i.e. the data loggings). We hope that both will allow us to have greater control over the expression and testing of fatigue-related team performance theories within a synthetic task environment. If we continue to specifically study team adaptation to fatigue (among other things), a reasonable manipulation that might increase such adaptation is to increase role cross-training (c.f. Cooke, Kiekel, Salas, Bowers, Stout, Cannon-Bowers, 2003), or perhaps force the issue a bit more by actually rotating participants through the roles as they do missions into the fatigue protocol.

In terms of data loggings, the most difficult issues seem to be measuring aspects of team coordination, not mission outcome or specific individual behaviors. Our challenge will be to log sequences of action across roles that can be easily correlated to infer aspects of team interaction in problem solving. For instance, we would have liked a more informative measure of jamming behavior (as moderated by Strike initially) in its use to support fighters (moderated by Sweep initially), so we could more easily tell when two assets were simultaneously directed against the same target (e.g. an enemy fighter). Future data logging should try to accommodate these kinds of team interactions more, and scenario design should be more forthcoming in requiring frequent need for such interaction.

Scenario design may be a bottleneck for these kinds of studies (only experts can write them well). However, we want to get enough insight on the process of scenario design to generate our future scenarios more algorithmically, yet remain unpredictable from the perspective of the participant. Other issues in scenario and synthetic task design reflect theoretical choices. For instance, if we want to study communication processes specifically (something being done elsewhere on the present data), our scenarios may need to entail some tunnel vision on the part of the roles, in the sense of giving each role only a partial world view that needs to be integrated (c.f. Brehmer, in press). Having the loggings record integrative communications (e.g. enforcing an email communication interface; having the task interface support schematized queries and answers) may go a long way toward solving these kinds of problems.

References

Anderson, J. R., (1987). Methodologies for studying human knowledge. *Behavioral and Brain Sciences*, 10, 467-505.

Barnes, C., Elliott, L. R., Coovert, M. D., & Harville, D. L. (2004). Effects of fatigue on simulation-based team decision making performance. *Ergometrika*, 4 (2 - 12). Online: http://www.ergometrika.org/volume4/BarnesEtAl.htm

Barnes, C. M., Petrov, P. V., Elliott, L. R., & Stoyen, A. (2002). Agent based simulation and support of C3 decisionmaking: Issues and opportunities. *Proceedings of the Eleventh Conference on Computer Generated Forces and Behavior Representation.* Orlando, FL.

Brehmer, B. (In Press). Some reflections on Microworlds research. In S. Schiflett, L. Elliott, E. Salas, & M. Coovert (Eds.) *Scaled Worlds: Development, Validation, and Applications*. Ashgate Publishing Limited, Surrey, England.

Brehmer, B., & Dorner, D. (1993). Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study. *Computers in Human Behaviour*, 9, 171-184.

Chaiken S. R., Elliott L., Dalrymple M., Coovert M., Riddle D., Gordon T., Hoffman K., Miles D., King T. (2001). Chaiken, S., Elliott, L. R., Dalrymple, M., & Schiflett, S. (2001). Weapons director intelligent agent-assist task: Procedure and findings for a validation study. *Proceedings of the 6th International Command and Control Research and Technology Symposium*, Annapolis, Maryland.

Cooke, N. J., Kiekel, P. A., Salas, E., Bowers, C., Stout, R., Cannon-Bowers, J. (2003). Measuring Team Knowledge A Window To The Cognitive Underpinnings Of Team Performance. *Group Dynamics: Theory, Research, and Practice*, 7 (3), 179-199.

Elliott, L. R., Dalrymple, M. A., Schiflett, S. G., & Miller, J. C. (In Press). Scaling Scenarios: Development and Application to C4ISR Sustained Operations Research. In S. Schiflett, L. Elliott, E. Salas, & M. Coovert (Eds.) *Scaled Worlds: Development, Validation, and Applications.* Ashgate Publishing Limited, Surrey, England.

Elliott, L. R., Coovert, M., Barnes, C., & Miller, J. C. (2003). Modeling performance in C4ISR sustained operations: A multi-level approach. *Proceedings of the 8th International Command and Control Research and Technology Symposium*. Washington, D.C.

Elliott, L. R., Barnes, C., Brown, L., Fischer, J., Miller, J. C., Dalrymple, M., Whitmore, J., & Cardenas, R. (2002). Investigation of complex C3 decisionmaking under sustained operations: Issues and analyses. *Proceedings of the 7th International Command and Control Research and Technology Symposium.* Quebec City, Canada.

Endsley, M. R. (1999). Situation awareness in aviation systems. In Garland, D. J. (Ed) & Wise, J. A. (Ed). *Handbook of aviation human factors. Human factors in transportation.* pp. 257-276 Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers.

Harville, D. L., Elliott, L. R., Dalrymple, M., Barnes, C., Miller, J. C., & Coovert, M. (2003, May). Communication and coordination in multi-operator mission performance over time: Effects of sleep deprivation on verbal and written communications. Paper presented at the International Conference on Naturalistic Decision Making, Pensacola Beach, FL.

Harville, D. L., Elliott, L. R., Coovert, M., Barnes, C., & Miller, J. C. (2003). Communication and decisionmaking in C4ISR sustained operations: An experimental approach. *Proceedings of the 8th International Command and Control Research and Technology Symposium.* Washington, D.C.

Hinsz, V.B. (2001). A groups-as-information-processors perspective for technological support of intellectual teamwork. In M.D. McNeese, E. Salas, & M.R. Endsley (Eds.), *New trends in collaborative activities: Understanding system dynamics in complex settings* (pp. 22-45). Santa Monica, CA: Human Factors & Ergonomics Society.

Hollenbeck, J., Ilgen, D., Sego, D., Hedlund, J., Major, D., & Phillips, J. (1995). Multilevel theory of team decision making: Decision performance in teams incorporating distributed expertise. *Journal of Applied Psychology*, 80(2), 292-316.

Hursh S. R. (1998). Modeling Sleep and Performance within the Integrated Unit Simulation System (IUSS). Technical Report Natick/TR-98/026L. Science and Technology Directorate, Natick Research, Development and Engineering Center, United States Army Soldier Systems Command, Natick, Massachusetts 01760-5020.

Klein, G. (2001). Features of team coordination. In M.D. McNeese, E. Salas, & M.R. Endsley (Eds.), *New trends in collaborative activities: Understanding System Dynamics In Complex Settings.* (pp. 68-95). Santa Monica, CA: Human Factors & Ergonomics Society.

Klein, G. (1996). The effects of acute stressors on decision making. In J. Driskell & E. Salas (Eds.) *Stress and human performance.* Mahwah, NJ: Lawrence Erlbaum Associates.

MacMillan, J., Paley, M.J., Levchuk, Y.N., Entin, E.E., Serfaty, D. & Freeman, J.T. (2002). Designing the Best Team for the Task: Optimal Organizational Structures for Military Missions. In Mike McNeese, Ed Salas, and Mica Endsley (editors), *New trends in collaborative activities: Understanding System Dynamics In Complex Settings.* (284-299). Santa Monica, CA: Human Factors & Ergonomics
Society.

Petrov, P., & Stoyen A. (2000). An Intelligent-Agent Based Decision Support System for a Complex Command and Control Application. *6th International Conference on Engineering of Complex Computer Systems* (ICECCS 2000), 11-15 September 2000, Tokyo, Japan. IEEE Computer Society 2000. 94-104.

Porter, O. L. H., Hollenbeck, J. R., Ilgen, D. R., Ellis, A. P. J., West, B. J., Moon, H. (2003). Backing Up Behaviors in Teams The Role of Personality and Legitimacy of Need. *Journal of Applied Psychology*, 88, 391-403.

Reeves, D., Winter, K., Kane, R., Elsmore, T., & Bleiberg, J. (2001). ANAM 2001 User's Manual. (Special Report NCRF-SR-2001-1). San Diego, CA: National Cognitive Recovery Foundation.

Zalesny, M. D., Salas, E., & Prince, C. (1995). Conceptual and measurement issues in coordination: Implications for team behavior and performance. In G. R. Ferris (Ed.), *Research in personnel and human resources management* (Vol. 13, pp. 81-115). Greenwich, CT: JAI Press.